图书情报工作
第63卷 第3期 2019年2月

基于用户兴趣度量的知识发现服务精准推荐*

丁梦晓[1] 毕强[1] 许鹏程[1] 李洁[1] 牟冬梅[2]
[1] 吉林大学管理学院 长春 130022 [2] 吉林大学公共卫生学院 长春 130021

摘要：［目的/意义］针对当前知识发现服务中存在的个性化程度不高和推荐效果不佳等问题，提出一种基于用户兴趣度量和内容分析的推荐算法。［方法/过程］文章通过特征词分布、LDA 主题分布、引文结构网络三个维度构建学术资源模型，并通过对用户行为的度量，计算用户对其浏览学术资源的兴趣度，结合学术资源模型构建用户兴趣模型。将用户兴趣模型与学术资源模型匹配，计算其相似度，得到用户对每条学术资源的兴趣值，最后将兴趣值最高的 TOP-N 学术资源推荐给用户。［结果/结论］通过实验检验算法的有效性和推荐准确率，结果显示，本文从实时动态度量兴趣的角度，提出的推荐算法能较好地预测用户兴趣，推荐效果显著，为实现发现服务精准推荐提供思路。

关键词：用户兴趣　内容分析　发现服务　精准推荐
分类号：G251
DOI：10.13266/j.issn.0252-3116.2019.03.003

我们已经由数字时代走向数据驱动时代，数据既是一种资产，又是一种资源。面对指数级增长、类型丰富的海量数据资源，如何对其进行有效利用，以实现面向用户的知识服务创新成为当下研究重点。知识发现服务作为知识服务的重要组成部分，是把资源和用户联系起来的重要环节。其中，如何准确把握用户兴趣偏好，预测用户需求，发现用户所需知识，并将其主动推荐给用户成为提升知识发现服务能力的突破点。然而，目前知识发现系统存在着灵活性不足、推荐结果不准和个性化服务程度不高等问题。同时，用户的使用环境和知识服务环境存在一定的融合鸿沟，融入用户环境的主动资源发现服务还有待加强。在数据驱动的大计算时代，“精准服务”是各行各业的发展方向，数字图书馆知识发现服务的发展思路也与之契合。精准推荐是提高知识发现服务质量的重要手段，在知识发现系统中融入精准推荐模式，通过分析用户的行为判断用户兴趣，聚合关联资源并充分利用新的技术手段为用户提供知识服务[1]，是发挥发现系统在数据资源方面的优势，满足用户个性化、精准化和知识化需求的重要手段。本文在分析发现系统用户行为集合的基础上，识别用户动态兴趣，构建出用户兴趣模型和资源内容模型，通过相关算法为用户进行精准的知识推荐。从而利用精准推荐技术来改变用户与知识发现系统的交互模式，协同系统精准推荐服务方式，为用户提供精准的推荐服务。

1 相关研究

自 2009 年网络资源发现系统 Summon 出现后，依托发现系统的知识发现服务已经发展了近十年，这期间对知识发现服务的研究主要集中在发现服务概念、功能分析、发现服务系统对比以及发现服务的应用上。对发现服务理论层面和功能应用层面的研究体现出学界和业界对提升知识发现服务质量的期望，但从精准推荐角度对发现服务进行研究的较少。

伴随着个性化成为知识服务的潮流，知识发现系统有必要为用户提供精准推荐来提升其服务质量。精准推荐是发现服务的关键一步，是用户对发现系统所提供服务的最深入感受。国内外对推荐服务的研究是从精准推荐的重要性和推荐算法两个方面展开。2004 年问世的 Google Scholar 基于元数据仓储的发现服务，在功能上支持相关文章推荐，引起热烈反响[2]，国内外学者相继投入到利用个性化推荐提升知识服务质量的

* 本文系国家自然科学基金面上项目“嵌入式知识服务驱动下的领域多维知识库构建”（项目编号：71573102）研究成果之一。
作者简介：丁梦晓（ORCID：0000-0002-6903-8360），硕士研究生；毕强（ORCID：0000-0001-7381-4986），教授，博士生导师；许鹏程（ORCID：0000-0001-5519-8550），硕士研究生；李洁（ORCID：0000-0002-3929-729X），博士研究生；牟冬梅（ORCID：0000-0003-0237-034X），教授，博士生导师，通讯作者，E-mail：moudm@jlu.edu.cn。
收稿日期：2018-07-06　修回日期：2018-10-10　本文起止页码：21-29　本文责任编辑：王传清

研究中。S. Q. Yang 和 K. Wagner[3]认为发现系统应具备推荐相关资源的服务功能,贯彻服务到人的理念为用户进行主题推送。在评价知识发现系统 EDS 和 Summon 时,美国林奇堡大学(Lynchburg College)图书馆[4]综合考虑了影响发现系统性能和质量的各种因素,将相似检索结果推荐(more like this)列为其重要指标。秦红[5]认为数字图书馆的资源发现系统应该具备个性化发现和自动化推荐等特征,在面向用户交互情境时为用户精准推送所需解决问题的知识。张钧[6]以用户的基本信息、行为信息等构建出用户画像,在此基础上预测用户的需求偏好及潜在知识需求,以此实现知识发现的个性化推荐匹配。如果说知识发现服务为用户发现了新知识和知识之间的隐性关联,那么精准推送服务则是为用户利用知识提供了一个专业化的获取与应用途径。通过精准推荐可以实现发现系统服务和用户的双赢,精准推荐是知识发现系统不可或缺的一部分,也是用户获得优质服务体验的途径。

精准推荐服务需要借助优质的推荐算法来实现,目前学术资源主流的推荐算法有基于内容的推荐算法和基于用户评分矩阵的协同过滤推荐算法。基于内容的推荐算法是通过对学术资源的内容特征进行提取,将内容相似的资源推荐给用户,推荐结果清晰直观。P. Guan 等[7]使用内容推荐的方法,通过合并元数据,如标题、关键词、摘要和引文来加强科学文献的语义信息,运用 TF-IDF 算法得到主题词权重向量,以此为基础构建用户兴趣模型,提高推荐的可理解性。但是,F. Ricci 等[8]指出基于内容的推荐方法只考虑了资源的内容特征,关注特征内容的意义性、结构性和易抽取性,没有充分考虑到用户的兴趣,没有完全达到个性化的目标。协同过滤推荐算法是通过用户行为数据和兴趣进行聚类实现,其基本原则是以相同的兴趣聚集用户,用户-项目评分相似的用户被认为具有相同的兴趣。当"邻居"用户浏览过某一项目而该用户没有浏览过时,则将该项目推荐给该用户。基于协同过滤的推荐策略需要借助用户评分信息来实现,在电子商务等领域推荐效果较好,应用比较广泛。然而协同过滤算法最棘手的问题是稀疏性和冷启动,即当一个系统的用户评分数据或涉及信息量较少时,推荐效果会大打折扣。第一个用户如何发现新物品,亦是有待解决的问题,而基于协同过滤的推荐算法无法针对新用户和新项目进行精准推荐,无法有效满足用户个性化需求。尤其是在数字图书馆领域,用户主动性相较于电子商务领域较差,当用户与数字图书馆交互的驱动力不足时,协同过滤算法的稀疏性和冷启动弊端更是被放大。

随着数据孤岛、信息过载、信息迷航等问题的凸显,传统推荐系统及其算法的顽疾尚未被解决,导致用户满意度降低,甚至出现用户流失的现象,制约了推荐服务的进一步推广和应用。传统推荐算法无法与用户兴趣和偏好变化的速度保持一致,对动态捕获用户潜在兴趣的推荐算法的研究就显得尤为重要。因此,信息服务提供商要充分考虑用户的动态兴趣,基于用户兴趣度量和内容分析来满足用户的动态知识需求,为用户提供更加精准的知识发现服务。鉴于此,本文将以精准推荐为目标,结合已有算法,基于知识发现系统海量资源数据和用户数据的优势,针对当前推荐存在的冷启动问题和用户兴趣转移问题开展相应的研究工作,系统地描述用户浏览学术资源时的隐式兴趣,识别出用户当前的兴趣状态,并提出具有一定创新性的基于用户兴趣模型的推荐算法。

2 资源内容度量

精准推荐的关键在于准确把握用户需求、兴趣或者偏好,深度挖掘资源内容特征,建立起用户与资源之间的联系,提供个性化知识推荐服务。因此,在精确推荐中,首先要建立学术资源模型,在此基础上与用户兴趣值相结合,建立用户兴趣模型,其次是如何使用用户兴趣模型进行精确推荐[9]。学术资源建模主要是对文本特征进行提取,特征词、主题词、引文等是学术资源的主要文本特征,通过提取学术资源的特征词分布、主题词分布、引文结构网络,从而构建学术资源模型,本文定义 M_d 表示学术资源模型,K_d 表示学术资源的特征词分布,T_d 表示学术资源的主题分布,C_d 表示学术资源引文,则学术资源模型表示为 $M_d = \{K_d, T_d, C_d\}$。

2.1 特征词分布

定义文档特征词集合 $K = \{K_{d1}, K_{d2}, \cdots, K_{dn}\}$,d 表示一个学术文本,文本特征词提取常用的方法是 TF-IDF 算法,即计算文档中词语的 TF-IDF 值,TF-IDF 值越大,则可作为文档的特征词。然而,传统的 TF-IDF 算法无法把握词语在文本集合中的分布比例量上的差异,这些差异正是表达文本内容的重要因素之一。因此,在 TF-IDF 的基础上引入信息增益的概念,改进传统的 TF-IDF 算法,特征词 K_{di} 的权重 W_{di} 的计算如公式(1)所示[10]:

$$W_{di} = \frac{TF_{di} \times \log(N \div n_d + 0.01) \times IG_d}{\sqrt{\sum_{i=1}^{t} (TF_{di} \times \log(N \div n_d + 0.01) \times IG_d)^2}}$$

公式(1)

其中，TF_{di}表示第i个特征词在学术文档d中出现的频率，N表示学术文档总数，n_d表示包含特征词i的学术资源数量。

而公式(1)中的IG_d为信息增益，表示词语的信息量，计算如公式(2)所示：

$IG_d = H(d) - H(d|i)$　　公式(2)

其中$H(d) = \sum(P(i) \times \log_2 P(i))$，$H(d|i) = -\sum(P(d|i) \times \log_2(P(d|i)))$，$P(i) = |wf(i)| / \sum|wf(i)|$　　公式(3)

$|wf(i)|$表示文档d中所有词的词频之和。

则向量$K_d = \{(K_{d1}, W_{d1}), (K_{d2}, W_{d2}), \cdots, (K_{dn}, W_{dn})\}$称为学术资源的特征词分布。

…

2.2　主题分布

定义主题分布$T = \{T_{d1}, T_{d2}, \cdots, T_{dn}\}$，d为一条学术资源，$T_{di}$表示文档di的主题分布概率，学术资源的主题分布采用LDA算法得到文档的主题和特征词的联合分布概率$p(w|d) = p(w|t) * p(t|d)$，利用Gibbs采样方法求解LDA模型的后验参数$P(T_{di}|d)$表示该学术资源d属于主题T_{di}的后验概率[11]，则向量$T_{di} = \{P(T_{d1}|d), P(T_{d2}|d), \cdots, P(T_{dn}|d)\}$称为学术资源的LDA主题分布。

2.3　引文结构

定义C_d表示学术资源的引文，则学术资源的引文集合用$C_d = \{C_{d1}, C_{d2}, C_{d3}, \cdots, C_{dn}\}$。科技文献之间的相互引证关系隐含了文献间的相似关系，通过引文关联可以找到一系列内容相关的文献，从而服务于推荐系统[12]。引文关联的建立可以根据科技文献科学的引证关系，运用图论理论构造引文图(citation graph)，一般以图G=(V,E)建模，顶点集V为信息对象集合，图上的任意点di∈V代表一篇引文。边集E表示顶点之间的关系，如果引文di引用了引文dj，则用边(di,dj)∈E来表示这个引用关系(见图1)。运用图论理论的方法挖掘隐含引文结构图中的顶点间的关系，利用图的拓扑结构信息计算引文结构相似度。

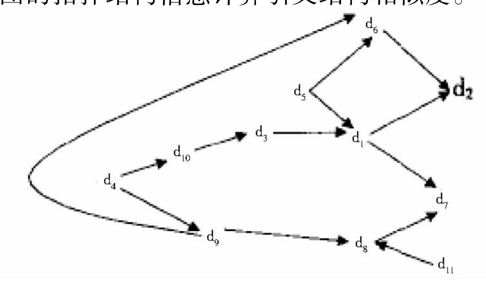


图1　引文结构图

3　用户兴趣度量

用户的兴趣偏好是推荐系统进行资源推荐的主要依据，用户兴趣度量的准确性直接影响知识发现服务精准推荐的质量。刘洪伟等[13]通过量化用户的动态隐性兴趣，为电子商务领域的个性化推荐服务提供帮助。曾子明等[14]从用户体验的视角判别用户兴趣的动态变化，在数字图书馆领域进行知识推荐服务。从相关研究中可以发现准确度量用户兴趣能够有效提高知识发现服务的推荐质量，产生更加精准的推荐效果。针对用户的兴趣分析与描述是知识发现服务实现精准推荐的基础，可通过建立用户兴趣模型来实现对用户兴趣的描述。用户兴趣模型描述了用户对资源信息的兴趣偏好，通过对用户兴趣的分析能够综合反映用户在一定时期内对资源信息的需求程度。

3.1　行为度量

用户兴趣可分为显性兴趣和隐性兴趣两种，显性兴趣是指用户通过主动的方式提供本人对知识需求的兴趣倾向，主要来源于用户进行正常注册时填写的个人信息所反映出的兴趣偏好；隐性兴趣指用户使用系统时产生的各种行为背后所隐含的兴趣偏好。因为显性兴趣通常比较稳定，且用户参与主动性较差，具有不准确性、不完全性和主观性等特点，无法反映用户的动态兴趣。而用户的隐性兴趣在数据采集过程中不需要用户的显式参与，只需在用户产生行为的同时记录数据即可，不影响用户浏览，因此本文采用隐性兴趣来动态地度量用户兴趣。当用户在浏览时，系统自动跟踪并记录服务器端的用户行为数据，基于行为数据计算出用户对页面内容的兴趣度，从而获取用户感兴趣的主题和内容。通过对用户行为数据的挖掘，得到的用户兴趣更加客观准确。

用户隐性兴趣的度量主要基于用户浏览行为，R. Krishnamoorthy[15]认为用户兴趣的度量是基于用户浏览行为的组合，并将浏览行为分为验证行为和致动行为两类。验证行为指可以用来判断用户兴趣有无的行为，如用户保存页面、打印页面和访问同一页面的次数等，这些行为展现出用户是否对浏览主题或者页面感兴趣，可以依此进行用户行为数据采集以判断用户兴趣程度；致动行为是验证行为的下一阶段，指可以判断出用户兴趣程度的行为，如用户在页面上的浏览时间、鼠标活动及键盘活动[16]等。L. Zheng等[17]重点对用户浏览时间与用户兴趣之间的关联关系进行综合分析，提出了通过用户浏览时间计算主题兴趣度的方法，

并证实该算法对计算用户兴趣合理准确，可有效用于个性化推荐。浏览时长是用户对浏览内容感兴趣程度在行为上的重要表现，用户的浏览时间越长，用户对页面内容的兴趣度就越高。当用户对打开页面及其内容感兴趣或者认为其有价值，用户会花费较长时间浏览页面。如果用户对于浏览的内容不感兴趣，则用户会快速关闭页面，点击下一个页面，重新寻找感兴趣的内容。用户浏览时间的影响因素主要有[18]：①用户对内容的关注度。用户对内容或主题的关注度越高，浏览时间越长。②页面内容量。页面的信息容量越大，用户花费在页面上的时间可能就越长。③用户理解能力。用户理解能力体现在当两个用户对同一个页面内容关注度一样时，用户理解能力越强，浏览所花费的时间就越短，因此应通过用户纵向对比度量用户对页面的兴趣度。由于用户间个体差异的存在，将用户的浏览绝对时间作为用户对某个页面兴趣度的测量依据有失偏颇，应当以同一用户浏览不同页面的相对时间的比值，同时考虑不同页面信息量的绝对比值作为衡量用户兴趣度的基准[15]。

另外考虑到用户学习兴趣转移的情况，以及为了反应用户近期的学习进度和兴趣，要选取用户一段时间 T 内浏览的学术文档以及其他交互行为进行度量。除了浏览时长和页面信息量两个度量指标，用户在浏览学术文档过程中，如果用户对某个学术资源特别感兴趣或者某个学术资源对用户特别有价值，用户就会进一步产生交互行为，如下载、收藏、分享等行为，而这些交互行为更加能体现用户的隐式兴趣，因此在度量兴趣时这些交互行为的权重应更高。基于以上考虑，得到用户对学术文档的兴趣度计算如公式(4)所示：

$$UI_i Interest = \frac{UI_i|time| * D_i content}{\sum_{i=1}^{n}(UI_i|time| * D_i content)} \qquad 公式(4)$$

if $|time| < Tmin$ $UI_i Interest = 0$

else {

　if download/collect/share

　$UI_i Interest = UI_i Interest + \delta$

　}

式中 $UI_i|time|$ 表示用户浏览学术文档的有效时间，$D_i content$ 表示学术文档的内容量，可以用学术文档的字节多少表示。T_{min} 是一个很小的值，旨在防止误点击，如果用户浏览学术文档 i 的时间小于 T_{min}，则认为是误点击，$UI_i Interest = 0$；如果大于等于 T_{min}，将通过公式(1)计算用户对 i 学术文档的兴趣度。如果用户对学术文档还有下载、收藏、分享等交互行为，认为用户对学术文档 i 的兴趣增加，则通过公式(1)计算得来的兴趣度 $UI_i Interest$ 增加 δ，δ 是调节参数，本文设置其值为 1。

3.2 用户兴趣模型

在学术资源模型的基础上，构建用户兴趣模型，定义 K_u 为用户兴趣的特征词向量，T_u 表示用户偏好主题分布，C_u 表示浏览文献的引文分布，则用户兴趣模型可表示为 $M_u = \{K_u, T_u, C_u\}$。

3.2.1 特征词偏好　知识发现系统中一项学术资源往往含有多个特征词，特征词可以对该资源内容进行简要概括和描述。令 $\{d_1, d_2, d_3, \cdots, d_i\}$ 表示某用户在一段时间 T 内浏览的所有学术资源的集合，通过分词工具和语料库，提取用户浏览的学术文档的特征词集合 $K_d = \{K_{d1}. K_{d2}, \cdots, K_{dn}\}$，则该用户的特征词偏好可用一个向量 $K_u = \{(K_{d1}, W_{d1}), (K_{d2}, W_{d2}), (K_{d3}, W_{d3}), \cdots, (K_{di}, W_{di})\}$ 描述。其中，K_{di} 表示第 i 个偏好特征词，W_{di} 为特征词 K_{di} 的权重。文本的特征词权重 W_{di} 的计算直接采用上文学术资源建模中 TF-IDF + IG 算法的计算结果。则有：

$$K_{ui}' = UI_i Interest * K_{ui} \qquad 公式(5)$$

其中，K_{ui} 则是学术资源的特征词分布，$UI_i Interest$ 表示用户对第 i 个特征词的兴趣度，K_{ui}' 表示用户浏览的学术资源新的特征词向量。

3.2.2 主题偏好　用户在一段时间 T 内浏览的某种学术资源的集合为 $\{d_1, d_2, d_3, \cdots, d_i\}$，用户的 LDA 主题偏好可用一个 N 维向量 $T_u = (T_{u1}, T_{u2}, T_{u3}, \cdots, T_{un})$ 描述。

$$T_{ui} = \frac{1}{N}\sum_{i=1}^{n} UI_i Interest \times T_{di} \qquad 公式(6)$$

其中，T_{di} 则是学术资源的主题概率分布，$UI_i Interest$ 表示用户对第 i 个主题的兴趣度，则 T_{ui}' 表示用户的兴趣主题分布。

3.2.3 引文分布　令 $\{d_1, d_2, d_3, \cdots, d_i\}$ 表示某用户在一段时间内阅读的某种学术资源的集合，建立引文关系图，则用户的引文集合用 $C_u = (C_{u1}, C_{u2}, C_{u3}, \cdots, C_{un})$ 表示。

4 精准推荐

4.1 相似度匹配

通过对学术资源文本特征进行提取，从特征词、主题词、引文三个维度建立学术资源模型 $M_d = \{K_d, T_d, C_d\}$，并结合用户兴趣度量，在学术资源模型的基础上建立用户兴趣模型 $M_u = \{K_u, T_u, C_u\}$。

使用 Jaccard 计算用户特征词偏 Ku 与学术资源特征词分布 Kd 的相似度，如公式(7)所示：

$$sim(i,j)=\frac{|K_u \cap K_d|}{|K_u \cup K_d|} \quad 公式(7)$$

使用余弦相似度算法计算用户主题偏好 Tu 与学术资源主题分布 Td 的相似度,如公式(8)所示:

$$sim(T_u,T_d)=\frac{T_u * T_d}{|T_u| \times |T_d|} \quad 公式(8)$$

根据引文结构图,运用 Sim Rank 算法,计算引文结构相似度,Sim Rank 递归定义相似度[19],常数 $c \in (0,1)$ 为阻尼因子,初始赋值如下:

$$\begin{cases} sim_0(C_{ui},C_{dj})=1,(if C_{ui}=C_{dj}) \\ sim_0(C_{ui},C_{dj})=0,(if C_{ui} \neq C_{dj}) \end{cases}$$

如果 $C_{ui} \neq C_{dj}$

$$sim_{l+1}(C_{ui},C_{dj})=\frac{c}{|I(C_{ui}I(C_{dj})|}\sum_{C'_{ui}\in I(C_{ui})}\sum_{C'_{dj}\in I(C_{dj})} sim_l(C'_{ui},C'_{dj}) \quad 公式(9)$$

如果 $C_{ui}=C_{dj}$, $sim_{l+1}(C_{ui},C_{dj})=1$

其中 I(C)表示指向 C 的临接点集合,如果 $I(C_{ui})$ 或 $I(C_{dj})$ 为空,则 $sim_{l+1}=(C_{ui},C_{dj})=0$。

公式(8)表示引文结构图中顶点 di 和 dj 之间的引文结构相似度,调用公式(8)递归 1 次,直到值收敛,最后的收敛值即为学术资源 C_{ui} 和 C_{dj} 的引文结构相似度。

定义用户的兴趣值 UID 为用户兴趣模型 Mu 和 Md 学术资源模型的相似度,计算如公式(10)所示:

$$UID=sim(M_u,M_d)=\frac{r_1 * sim(K_u K_d)+r_2 * sim(T_u,T_d)+r_3 * sim(C_u,C_d)}{\sqrt{r_1^2+r_2^2+r_3^2}} \quad 公式(10)$$

其中 $r_1+r_2+r_3=1$,具体权重根据实验训练分配。将用户兴趣值 UID 最高的 TOP-N 推荐给用户。

4.2 推荐算法流程

如图 2 所示,本文提出的精准推荐算法具体流程如下:①通过网络爬虫工具获取学术资源;②提取学术资源的信息(资源 ID、标题、摘要、关键词、引文等),并建立引文结构网络图;③对提取的学术资源的信息进行预处理(分词、去停用词等);④计算每条学术资源的特征词分布、LDA 主题分布以及引文相似度,构建学术资源模型;⑤基于 Web 日志记录用户行为(浏览时间、下载、转发、收藏等),计算用户浏览过的学术资源的兴趣度;⑥基于用户兴趣度和学术资源模型,构建用户兴趣模型;⑦计算用户兴趣模型与学术资源模型相似度,得到用户对每条学术资源兴趣值[20];⑧将兴趣值最高的 TOP-N 学术资源推荐给用户。

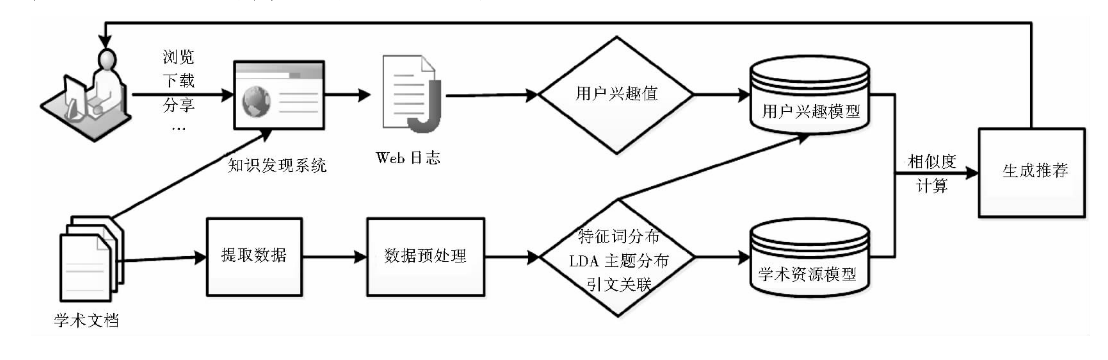


图 2 推荐流程图

5 实验

5.1 数据采集及处理

本实验的数据源来源于中国学术期刊(网络版)CAJ-N 数据库,选择图书情报与数字图书馆专题领域中文期刊(55)2007-2018 年的论文作为实验数据,剔除专题序、会议通知、不完整论文等,总共获取有效论文 10 227 篇,爬取论文的标题、摘要、关键词、引文。在进行分词前将实验的论文的关键词、《图书馆学情报学大辞典》和《汉语主题词表》作为分词词典导入中国科学院 NLPIR 汉语分词系统,并建立同义词表和停用词表,以改进分词效果。对学术资源的标题、摘要、关键词进行分词处理,并去停用词。随后对特征词进行词频统计,计算特征词的 TF-IDF 值,筛选出 TF-IDF 值排名前 5 的名词或动词作为特征词,将文本向量化,表示为文档-特征词矩阵,从而构建出学术资源的特征词分布模型。在 LDA 建模过程中,利用 MCMC 方法中的 Gibbs 采样法进行参数估计,其中主题数 K=50,设置文档-主题超参数 $\alpha=0.2$,主题-词项分布的参数 $\beta=0.01$,Gibbs 采样迭代次数设置为 1 000 次。通过 LDA-Gibbs 模型训练计算,我们得到 10 227 篇文献的文档-主题分布和 K 个主题的词项分布,部分主题词及关键词分布如表 1 所示:

表 1 部分主题词和关键词分布

用户研究	信息服务	资源组织	知识管理	评价研究	资源共享	知识产权	企业情报
用户	服务	数据	图书	评价	建设	知识	企业
需求	信息	系统	知识管理	标准	资源	情报	竞争
信息	个性化	检索	图书馆	体系	图书馆	保护	情报
系统	用户	语义	服务	指标	共享	知识产权	分析
交互	数字	模型	组织	评估	特色	科学	战略
设计	数字图书馆	本体	信息	模型	高校	研究	决策
调查	模式	推荐	理论	质量	文献	网站	创新
因素	需求	元数据	创新	信息	联盟	版权	产业
模型	咨询	技术	策略	绩效	服务	科技	环境
数字图书馆	读者	结构	机制	服务	共建	软件	技术
行为	方式	关联	企业	维度	机制	法律	研究
问卷	智能	资源	能力	可用性	平台	制度	市场
界面	主动	用户	体系	理论	馆藏	中国	管理
感知	质量	模块	资源	规范	模式	利益	智库
情境	推送	算法	社区	风险	合作	许可	风险

利用 UCINET 软件，构建实验论文的引文网络如图 3 所示。其中，节点表示科技文献，节点间连线的方向指明了文献间的引用与被引用关系，通过引证关系揭示文献之间的关联，通过 SimRank 算法计算顶点的相似度。



图 3 引文网络图

5.2 实验设置

为验证所提出的精准推荐算法的准确性，本研究邀请 30 位图情专业的学生作为实验对象，每位用户根据自己的兴趣或者任务，在图书情报领域目录下进行至少 20 次检索行为，以保证获取充分的用户行为数据。用户的浏览时间、检索、收藏、下载、转发、拖动滚动条、翻页等多种行为数据通过嵌入 JavaScript 代码进行获取。实验过程中，将整个用户行为数据集分为两部分，80% 作为训练集，以产生用户兴趣模型，保留 20% 作为测试集用于验证算法推荐效果。用户兴趣模型构建中结合已构建好的资源内容模型，通过本文提出的算法计算每位实验对象的 UID 兴趣值，其中设置 $r_1 = 0.3$，$r_2 = 0.4$，$r_3 = 0.2$，分别将用户兴趣值最高的 TOP－5\TOP－10\TOP－15\TOP20 的资源推荐给用户，共推荐 10 次，每次推荐后用户对自己感兴趣的资源进行访问。

5.3 结果评价

5.3.1 推荐效果评价指标　为评测构建的模型的推荐效果，本文选取了准确率(precision)、召回率(recall)、F 值三个评价指标对推荐结果进行评估。计算如公式(10)所示：

$$P = \frac{A}{A+B} \times 100\%$$

$$R = \frac{A}{A+C} \times 100\% \qquad \text{公式(10)}$$

$$F = \frac{2 \times P \times R}{P+R}$$

其中，A 表示推荐的感兴趣的资源数量，B 表示推荐的不感兴趣的资源数量，C 表示未推荐的感兴趣的资源数量。

5.3.2 对比实验调试　本文实验选取基于内容(LDA 主题模型)的推荐算法和基于用户的协同过滤推荐算法进行对比。在对比实验中，利用 LDA 主题模型进行建模时，需要设定主题个数 K 的大小，表 2 显示当给每个用户推荐学术资源数为 20 时，不同的 K 值对准确率、召回率和 F 值的影响，可以看出 K 为 20 的时候是最佳值。

表 2　不同的主题个数对应的准确率、召回率和 F 值

主题个数	10	15	20	25	30
召回率	0.125	0.187	0.246	0.174	0.136
准确率	0.386	0.405	0.427	0.352	0.291
F 值	0.189	0.256	0.312	0.233	0.185

基于用户的协同过滤算法进行推荐时，设置的最近邻居的个数不同，推荐效果会有所差异。表 3 显示当给用户推荐的学术资源数为 20 时，不同最近邻居个数 h 值对准确率、召回率和 F 值的影响，可以看出最近邻居个数为 30 的时候推荐效果最佳。

表 3　不同的最近邻居个数下对应的准确率、召回率和 F 值

最近邻居数	10	20	30	40	50
准确率	0.349	0.453	0.526	0.427	0.324
召回率	0.193	0.221	0.27	0.209	0.198
F 值	0.249	0.297	0.357	0.281	0.246

5.3.3　结果对比　根据对比实验的调试结果，设置主题个数为 20、最近邻居数为 30 时，基于内容的推荐算法和基于用户的协同过滤算法的推荐效果最好。在该实验条件下分别计算三种算法不同推荐个数下的准确率、召回率和 F 值以及实验的平均准确度（precision），实验结果如图 4－图 7 所示。

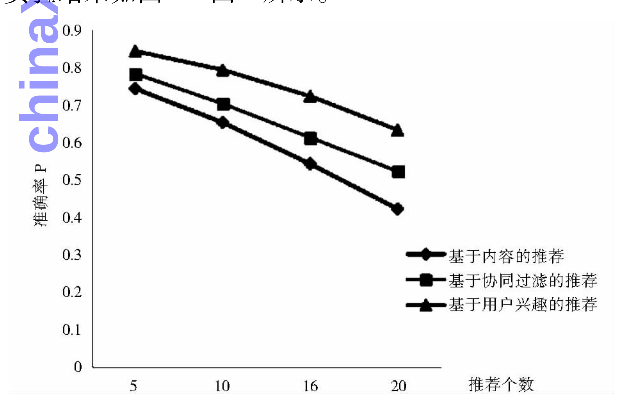


图 4　不同推荐个数下准确率比较（P 值）

实验结果显示，当推荐个数从 5 依次上升到 20 时，各种方法的准确率依次降低，召回率和 F-measure 值依次上升。当推荐个数相同时，本文提出的基于用户兴趣度量的推荐算法的准确率和 F 值都是最高的，其推荐效果最好，其次是协同过滤算法，最后是基于内容的推荐算法。综合整个实验，基于用户兴趣度量的协同过滤算法的平均准确度比基于协同过滤算法的推

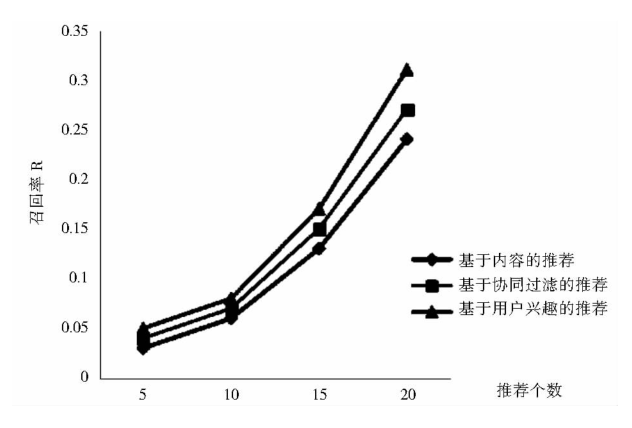


图 5　不同推荐个数下召回率比较（R 值）

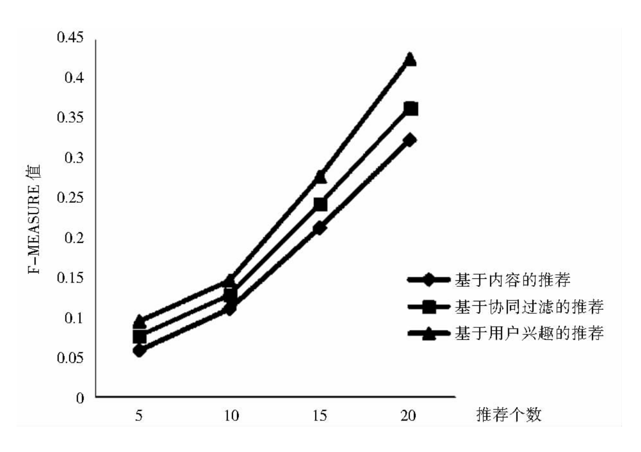


图 6　不同推荐个数下 F 值比较

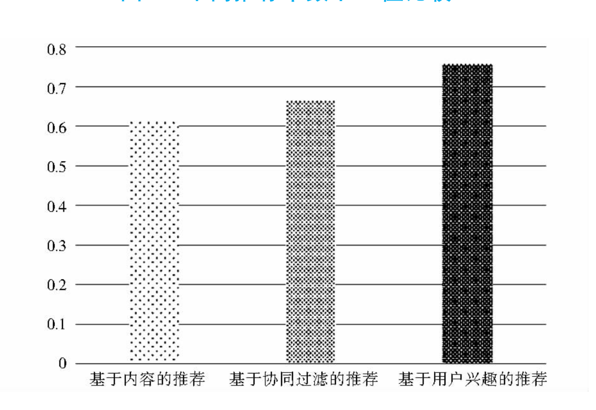


图 7　平均推荐准确度

荐准确度提升 14%，比基于内容的推荐算法的推荐准确度提升 23%。由此可见，考虑了用户行为和引文关联，使得本文提出的算法更能预测用户兴趣，其推荐效果也更好。

6　结语

知识发现系统以其丰富的资源数据和用户数据为精准推荐服务提供了数据基础。通过对数据资源进行碎片化处理、细粒度挖掘和分析，发现系统可以深层次呈现资源的内容特征，揭示其语义关系，建立引文关

联,实现资源的深度聚合,为用户发现资源间的隐含联系,揭示新的知识模式,提供精细化的知识发现服务。随着用户需求的碎片化、精细化和个性化,发现系统需要充分利用用户行为数据,度量用户兴趣、了解用户需求,为用户提供精准的知识推荐服务,提升用户交互体验,满足用户知识需求,促进知识价值的倍增。精准推荐是增强数字图书馆知识发现服务能力的重要功能,为数字图书馆知识发现服务带来了创新生长点。本文从用户兴趣的角度出发,通过从特征词、主题词和引文三个维度提取资源内容特征,建立学术资源模型,并通过对用户兴趣的度量,构建用户兴趣模型,运用相似性算法对知识发现服务的精准推荐进行优化,再通过实证检验算法的可行性,与传统基于内容推荐算法和基于协同过滤的推荐算法进行对比。本文提出的推荐算法有以下三方面优势:①考虑引文关联,更加科学地揭示学术资源间的内在联系。②引入了用户行为集合,对用户的兴趣偏好程度进行分析,推荐结果更加准确、客观。③当用户兴趣发生改变时,推荐算法可以通过捕捉用户近期兴趣的改变而推荐更为适合的信息。本文的推荐算法可以实时把握用户兴趣,为用户进行精准推荐,提升发现系统知识服务能力,改进用户的使用体验。本文也存在一些不足之处,如算法步骤繁琐,计算量较大,实验样本和时间存在局限,实验过程中部分环节需要人工控制,带有一定的主观性等。因此在下一步研究中,笔者将进一步提升算法性能,简化算法步骤,提升算法的适用性,增强推荐结果的准确性。

参考文献:

[1] 毕强,刘健. 基于领域本体的数字文献资源聚合及服务推荐方法研究[J]. 情报学报,2017,36(5):452 -460.
[2] WALTERS W H. Google Scholar coverage of a multidisciplinary field[J]. Information processing & management, 2007, 43(4):1121 -1132.
[3] YANG S Q, WAGNER K. Evaluating and comparing discovery tools: how close are we towards next generation catalog? [J]. Library hi tech, 2010, 28(4):690 -709.
[4] MICHAEL G. The evaluation of discovery services at Lynchburg College: 2009 - 2010[J]. College & undergraduate libraries, 2012, 19(2 -4):387 -397.
[5] 秦红. 普适计算环境中的数字资源感知服务框架探讨[J]. 图书情报工作,2014,58(5):13 -16,21.
[6] 张钧. 基于用户画像的图书馆知识发现服务研究[J]. 图书与情报,2017(6):60 -63.
[7] GUAN P, WANG Y F. Personalized scientific literature recommendation based on user's research interest[C]// International conference on natural computation, Fuzzy systems and knowledge discovery. Changsha:IEEE, 2016:1273 -1277.
[8] RICCI F, ROKACH L, SHAPIRA B, et al. Recommender systems handbook[M]. New York:Springer, 2011.
[9] RAZMERITA L. An ontology-based framework for modeling user behavior-a case study in knowledge management[J]. IEEE transactions on systems, man, and cybernetics - part A: systems and humans, 2011,41(4):772 -783.
[10] 李学明,李海瑞,薛亮,等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程,2012,38(8):37 -40.
[11] 王振振,何明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学,2013,40(12):229 -232.
[12] 王传清,毕强. 超网络视域下的数字资源深度聚合研究[J]. 情报学报,2015(1):4 -13.
[13] 刘洪伟,高鸿铭,陈丽,等. 基于用户浏览行为的兴趣识别管理模型[J]. 数据分析与知识发现,2018(2):74 -85.
[14] 曾子明,金鹏. 基于用户兴趣变化的数字图书馆知识推荐服务研究[J]. 图书馆论坛,2016,36(1):94 -99.
[15] KRISHNAMOORTHY R, SUNEETHA K R. User interest estimation using behavior monitoring measure [J]. Transplantation, 2013, 78(2):651 -652.
[16] CLAYPOOL M, BROWN D, LE P, et al. Inferring user interest [J]. IEEE internet computing, 2001, 5(6):32 -39.
[17] ZHENG L, CUI S, YUE D, et al. User interest modeling based on browsing behavior [C]// International conference on advanced computer theory and engineering. Chengdu:IEEE, 2010:V5 -455 - V5 -458.
[18] 张海鹏. 基于 Web 日志挖掘的个性化推荐研究[D]. 重庆:重庆大学,2007.
[19] JEH G, WIDOM J. SimRank: a measure of structural-context similarity[C]// Eighth ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton:ACM, 2002:538 -543.
[20] 尹丽玲,刘柏嵩,王洋洋. 跨类型的学术资源优质推荐算法研究[J]. 情报学报,2017,36(7):715 -722.

作者贡献说明:

丁梦晓:设计研究方案,撰写论文;
毕强:提出研究思路,修改论文;
许鹏程:数据采集及实验;
李洁:完善研究思路,修改论文;
牟冬梅:完善研究思路,修改论文。

Research on Precise Recommendation of Knowledge Discovery Services Based on Users Interests

Ding Mengxiao[1] Bi Qiang[1] Xu Pengcheng[1] Li Jie[1] Mu Dongmei[2]

[1] School of Management, Jilin University, Changchun 130022

[2] School of Public Health, Jilin University, Changchun 130021

Abstract: [**Purpose/significance**] This paper proposes a recommendation algorithm based on user interest metrics and content analysis for the current issues of low personalization and poor recommendation in knowledge discovery services. [**Method/process**] Through characteristic word distribution, LDA topic distribution and citation association, this paper constructs the academic resource model. Through the measurement of user behavior (browsing time, downloading, forwarding, collecting, etc.), the user's interest in browsing academic resources can be calculated, and the user interest model is constructed. Matching the user interest model with the academic resource model and calculating its similarity, the user's interest value for each academic resource can be obtained. Finally, the TOP-N academic resources with the highest interest value can be recommended to the user. [**Result/conclusion**] The paper tests the effectiveness of the algorithm and the accuracy of the recommendation through experiments. From the experimental results, we can show that the recommendation algorithm can predict the user's interest better and the recommendation effect is significant, simultaneously providing ideas for precise recommendation of discovery services.

Keywords: user interest content analysis discovery service precise recommendation